Introduction/Background:

The proliferation of phishing attacks poses a significant threat to digital security, with attackers leveraging increasingly sophisticated tactics to evade detection. The advent of artificial intelligence (AI) has raised concerns that AI-powered phishing attacks will become even more targeted, convincing, and difficult to detect. New public tools such as OpenAI's ChatGPT has allowed cybercriminals to generate more targeted and human-like phishing emails on a large scale. Since the fourth quarter of 2022 (the public release of ChatGPT) there has been a 1,265% rise of phishing emails according to cybersecurity firm SlashNet[i].

Local Large Language Models (LLMs) will likely become the most popular approach to constructing AI phishing messages as cloud providers such as OpenAI and Google (which both offer free LLM access in their cloud infrastructure) scan queries for malicious intent[ii].

Spear phishing in particular will become a larger threat with the increased access to AI models. Spear phishing is a technique of phishing which targets the victim specifically, often including their name, username, and other information that the victim often assumes that an attacker doesn't know, leading to a increased rate of success for this type of attack. AI makes writing targeted messages such as these much easier to create on a larger scale[iii].

Research Objectives:

- Create a dataset of AI-generated phishing messages for researchers to analyze and use to create systems which detect and counter AI generated phishing attempts.
- Perform a preliminary analysis of said dataset to determine characteristics which may be used to detect AI generated phishing attempts.

Methodology:

To generate enough data for reliable analysis, we will use UWEC's HPC center to run local LLMs and programmatically instruct them to create targeted phishing messages and store them in the corpus.

We will follow the 'best practices' for creating phishing messages, those being (roughly)[iv]

- Sense of urgency
- Sense of legitimacy
- Personalization to the target

Once the dataset is generated, we will conduct a basic analysis of its characteristics by examining linguistic features such as grammar, syntax, and semantics of text-based phishing emails, to determine with basic confidence whether existing language solutions such as NLP (Natural Language Processing) can accurately detect AI phishing attempts, as well as provide recommendations for avenues of research into other characteristics for detection.

Finally, we will release the dataset for researchers to access and use, likely through a platform commonly used for Machine Learning or Artificial Intelligence training data storage and sharing, such as Kaggle[v], HuggingFace[vi], or Github[vii].

[i] The State of Phishing 2023. (2023). In *slashnext.com*. Slashnext. https://slashnext.com/wp-content/uploads/2023/10/SlashNext-The-State-of-Phishing-Report-2023.pdf

[ii] Shevlane, T. (2023, May 25). *An early warning system for novel AI risks*. Google DeepMind. https://deepmind.google/discover/blog/an-early-warning-system-for-novel-ai-risks/

[iii] Microsoft. (2023, July 14). *How AI is changing phishing scams*. Microsoft 365. https://www.microsoft.com/en-us/microsoft-365-life-hacks/privacy-and-safety/how-ai-changing-phishing-scams

[iv] *Phishing and Social Engineering | Account Security and Fraud Claims*. (n.d.). @Verizon. https://www.verizon.com/about/account-security/phishing

[v] kaggle.com

[vi] huggingface.co

[vii] github.com